Morphological Productivity in the Lexicon

Onur T. Şehitoğlu and H. Cem Bozsahin
Laboratory for the Computational Studies of Language
Department of Computer Engineering
Middle East Technical University, Ankara, Turkey
{onur,bozsahin}@LcsL.metu.edu.tr

Abstract

In this paper we outline a lexical organization for Turkish that makes use of lexical rules for inflections, derivations, and lexical category changes to control the proliferation of lexical entries. Lexical rules handle changes in grammatical roles, enforce type constraints, and control the mapping of subcategorization frames in valency changing operations. A lexical inheritance hierarchy facilitates the enforcement of type constraints. Semantic compositions in inflections and derivations are constrained by the properties of the terms and predicates.

The design has been tested as part of a HPSG grammar for Turkish. In terms of performance, run-time execution of the rules seems to be a far better alternative than pre-compilation. The latter causes exponential growth in the lexicon due to intensive use of inflections and derivations in Turkish.

1   Introduction

Languages like Finnish, Hungarian, and Turkish have relatively rich morphology which governs grammatical functions often delegated to syntax in languages such as English. Prominence of morphology puts a greater demand on the information in the lexicon, which may grow to an unmanageable size due to heavy use of inflections and derivations. In Turkish, for instance, the nominal paradigm has three affixes (number, case, relativizer), and the verbal paradigm has eight (for voice, tense, person, aspect, and mood). Generating the full paradigm for a nominal and a verbal root requires 2^3 and 2^8 entries in the lexicon, respectively. The problem is further complicated by the rich inventory of derivational affixes for both paradigms, as exemplified in 1. Hankamer [**?**] argues convincingly that full listing of every word form in the lexicon is untenable for agglutinative languages.

(1) *Yaz-ıcı-lar-a*   *gör-ev-ler-i*   *bil-dir-il-me-miş-ti*
    write-VtoN-PLU-DAT able-VtoN-PLU-ACC know-CAUS-PASS-NEG-ASP-TENSE
    'The clerks have not been informed of their duties'

Handling inflections and derivations with lexical rules opens us possibilities for encoding semantic and grammatical changes in the lexicon as well. For instance, a causative suffix will demote an agent to a patient or a recipient, and it will add a new grammatical role for the causer (the new agent). A locative case suffix will mark a NP as an adjunct, which can no longer satisfy subcategorization requirements of the verbs or postpositions. We elaborate on the consequences of these phenomena in section 3.

Another source for economy of representation can be seen in example (2), where attributive adjectives are used as nouns in 2b and 2d. One solution to this problem is syntactic underspecification, e.g., grouping the nouns and adjectives under a single lexical category.[1] An alternative is to use a lexical rule for differentiating predicate and term reading of the lexical entry.

(2) a. *kuru yaprak*
        dry   leaf
        'dry leaf'

    b. *meyve kuru-su*
        fruit   dry-POSS
        'dried fruit'

    c. *yaş-lı   hanım*
        age-ADJ lady
        'old lady'

    d. *bütün yaş-lı-lar*
        all   age-ADJ-PLU
        'all elderly'

In what follows, we will describe different kinds of lexical rules for type constraints, and handling changes in grammatical roles or subcategorization requirements. We also discuss processing issues such as run-time generation versus pre-compiling of word forms.

2 Morphology-syntax Interface

Modelling inflections, derivations, and the corresponding phonological alternations via lexical rules amounts to the lexicalization of morphology. The alternatives to this approach (for Turkish) have also been explored, e.g., the modularization of syntax and morphology by keeping them (and their lexicons) as separate systems that communicate with each other [?], or integrating morphology, syntax and semantics, thus treating morphotactics in the same manner as syntax with respect to semantic composition [?]. From a computational point of view, the modular approach has efficient lexical access since lexical search is performed on root forms, and bound morphemes are not considered lexical items. In the integrated (multi-dimensional) approach, the lexicon contains free and bound morphemes; they have complete syntactic and semantic specifications. Some of the inflections, e.g. person and number, do not have any contribution to semantics, hence their semantic form (or LF) is that of identity. Some inflections, such as case and causative affixes, compose semantic form of the stem (LF_s) with that of the affix. LF_s can be turned into (*cause* x LF_s) for causatives where x is the new argument introduced by the causative affix.[2] Similar arguments can be made for the semantic contribution of adjunct case markers.

The lexical approach to morphology presented here is a mid-point in the design of the morphology-syntax interface. In this view, morphology is not isolated from syntax, but, similar to the modular organization, bound morphemes are not considered lexical items. They can be attached to stems

---

[1]In fact, traditional Turkish grammar books such as [?] collectively call them "substantives."

[2]cf. example 9

via lexical rules. This implies that lexical rules are responsible for semantic composition and for the changes in syntactic requirements. This view also represents a middle ground in the complexity of lexical structures.

Keeping morphology and syntax entirely separate forces one to stipulate different scopes for affixes. For instance, the adverbial suffix -*ken* and the adjectival -*lu* might have phrasal (3a and 3c) or lexical scope (3b and 3d). Multi-dimensional approach allows affixes to 'pick out' different scopes in mixed morphological and syntactic composition. The lexical approach can accomodate both readings, provided that lexical rules are invoked with relevant syntactic information, e.g., valency of the verb. Morphologically ambiguous cases such as 4 are handled by multiple instantiations of the lexical rules.

(3) a. *Çocuk top-a*   *[kaleci-ye*   *bakar]-ken vurdu*
     child   ball-DAT goalkeeper-DAT look-ADV hit
     'The child hit the ball facing the goalkeeper.'

    b. *Çocuklar [yürür]-ken taş*   *toplamışlar*
     children   walk-ADV   stone picked
     'The children had picked stones while walking.'

    c. *[Uzun kol]-lu*   *gömlek*
     long   sleeve-ADJ shirt
     'shirt with long sleeves'

    d. *Uzun [çiçek]-li*   *gömlek*
     long   flower-ADJ shirt
     'long shirt with flower patterns'

(4) a. *kalem-ler-i*   b. *kalem-ler-i*   c. *kalem-leri*
     pencil-PLU-ACC   pencil-PLU-POSS.3SG   pencil-POSS.3PL
     'the pencils (=OBJ)'   'his/her pencils'   'their pencils'

It is too early to evaluate the advantages and disadvantages of these approaches in terms of competence grammars and performance issues. But the choice of the strategy also affects the design of lexical organization. For instance, if inflections and derivations are handled by lexical rules, the morphological features need not be kept in the lexicon, since the lexical rules will reflect the changes in syntactic and semantic requirements coming from morphology. If morphology is treated almost like syntax, lexical knowledge should contain richer morphological information, including a semantic representation for bound forms (affixes), information about boundedness/freeness of morphemes, and the type of attachment (e.g., affixation, cliticization, syntactic concatenation) [**?**, **?**]. This will enable the system to rule out, for instance, affixation of two free forms, or impose selectional restrictions on the stems of affixes.

In this study, a lexical inheritance hierarchy is used in conjunction with the lexical rules to obtain type constraints and feature structures for free forms (words); bound forms are not part of the lexicon. The hierarchy is given in Figure 1.

This tree is part of a greater hierarchy which includes inheritance information for words and phrases. We make use of the inheritance and type-checking mechanism of ALE [**?**] to impose type-specific constraints on words. Words are distinguished from phrases by disallowing any kind of
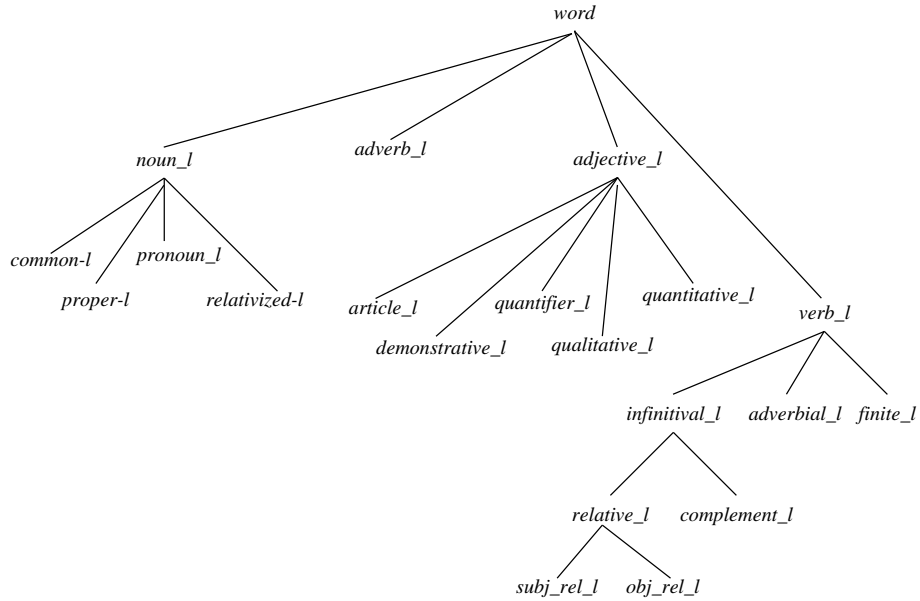
Figure 1: Lexical hierarchy

gapping below the word level in the tree. Designating a lexical item as one of the subtypes in the hierarchy will apply all the constraints and incorporate the feature structures of the supertypes along the path to *word*. For instance, a qualitative adjective (e.g., *rahat*=comfortable) is distinguished from a quantitative one (e.g., *çift*=double) by its choice of modifiers; the latter does not allow intensifiers (5).

(5) a. *çok rahat koltuk*
very comfortable couch
'very comfortable couch'

b. ** çok çift koltuk*

c. *rahat çift koltuk*
comfortable double couch
'comfortable twin couch'

The fragments[3] of the type constraints for these subtypes are given in Figure 2. The controlled use of type constraints at different levels of the lexical hierarchy eliminate the need to enumerate type-specific lexical rules to achieve the same effect.

3 Types of lexical rules

Inflections: Lexical rules for inflections can check morphotactic constraints for proper ordering of morphemes. More importantly, they should reflect the grammatical or semantic requirements imposed by inflections. For instance, the locative case suffix in Turkish also marks an NP as adjunct (6).

---

[3]We use HPSG style feature structures and signatures in our descriptions. See Pollard and Sag [**?**].

$$
\begin{bmatrix}
\textit{quantitative-l} \\
\text{SYNSEM}\,|\,\text{LOCAL}\,|\,\text{CAT}\,|\,\text{HEAD} \quad \boxed{1}
\end{bmatrix} \Rightarrow
$$

$$
\boxed{1}
\begin{bmatrix}
\textit{quantitative-adj} \\
\text{MOD}
\begin{bmatrix}
\text{MODSYN}
\begin{bmatrix}
\text{LOCAL}\,|\,\text{CAT}
\begin{bmatrix}
\text{HEAD} \quad \textit{common} \\
\text{ADJUNCTS}
\begin{bmatrix}
\text{QUANT} & - \\
\text{QUANT-ADJ} & - \\
\text{QUAL-ADJ} & \boxed{2} \\
\text{NON-REF} & \boxed{3}
\end{bmatrix}
\end{bmatrix}
\end{bmatrix} \\
\text{MODADJ}
\begin{bmatrix}
\text{QUANT} & - \\
\text{QUANT-ADJ} & + \\
\text{QUAL-ADJ} & \boxed{2} \\
\text{NON-REF} & \boxed{3}
\end{bmatrix}
\end{bmatrix}
\end{bmatrix}
$$

$$
\begin{bmatrix}
\textit{qualitative-l} \\
\text{SYNSEM}\,|\,\text{LOCAL}\,|\,\text{CAT}\,|\,\text{HEAD} \quad \boxed{1}
\end{bmatrix} \Rightarrow
$$

$$
\boxed{1}
\begin{bmatrix}
\textit{qualitative-adj} \\
\text{MOD}
\begin{bmatrix}
\text{MODSYN}
\begin{bmatrix}
\text{LOCAL}\,|\,\text{CAT}
\begin{bmatrix}
\text{HEAD} \quad \textit{common} \\
\text{ADJUNCTS}
\begin{bmatrix}
\text{QUANT} & \boxed{2} \\
\text{QUANT-ADJ} & \boxed{3} \\
\text{NON-REF} & \boxed{4}
\end{bmatrix}
\end{bmatrix}
\end{bmatrix} \\
\text{MODADJ}
\begin{bmatrix}
\text{QUANT} & \boxed{2} \\
\text{QUANT-ADJ} & \boxed{3} \\
\text{QUAL-ADJ} & + \\
\text{NON-REF} & \boxed{4}
\end{bmatrix}
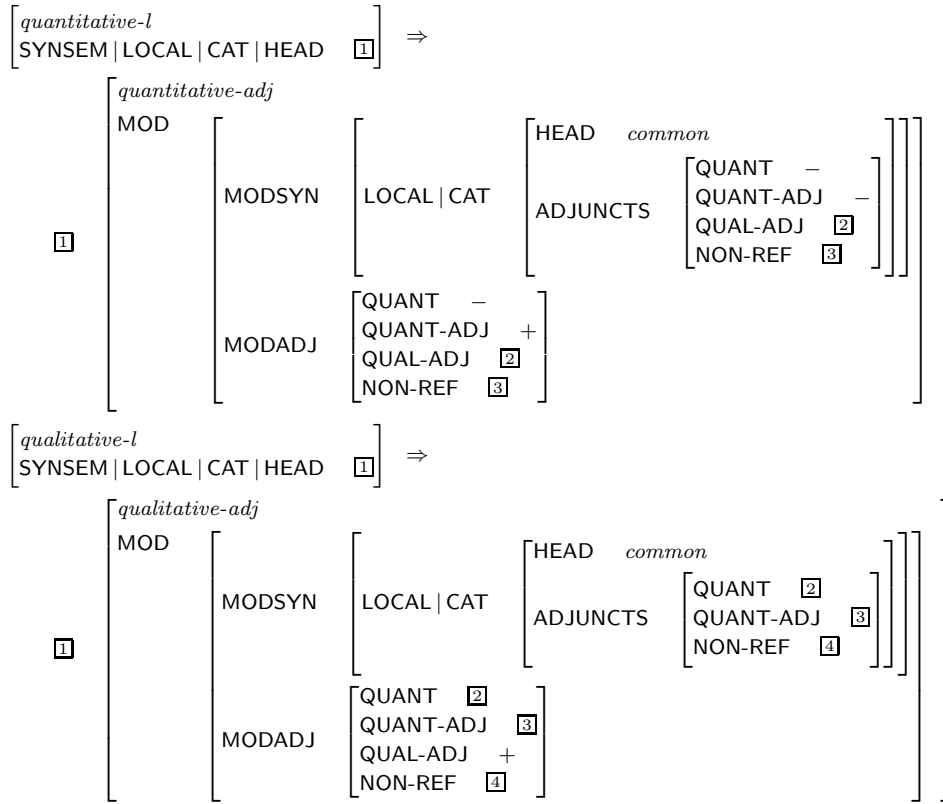\end{bmatrix}
\end{bmatrix}
$$

Figure 2: Type constraints for words and some subtypes.

(6) *Adam araba-da uyu-du* .
    man   car-LOC sleep-TENSE.3SG
    'The man slept in the car'

The lexical rule for locative case is given (in ALE notation) in Figure 3. This rule is applied when the locative suffix is attached to a nominal stem. The head of the NP is marked with the locative case, and the type of NP is changed to an adjunct. This is achieved by modifying the head feature MOD: While the nominative marked noun has null value, a MODSYN value with verbal head is introduced in the head feature of the locative noun. This will allow the locative marked noun to modify a verb. Thus, it cannot satisfy the subcategorization requirements of verbs or postpositions. This issue is critical for parsing relatively free word-order languages where grammatical relations are often indicated by overt case marking rather than structural position. Figure 3 also shows the derivation of the semantic representation for the case marked NP; $at(x,y)$ is a second-order predicate that holds between a term x and a predicate y. This predicate is inserted into the set of restrictions for the noun. Although this method is not generative in the sense of [**?**], it allows semantic composition in the lexicon.

Derivations: Denominal verbs, deverbal nouns, and part of speech changes can be modelled respectively by adding subcategorization frames, discharging subcategorization frames, and type coercions, via lexical rules. The most difficult issue in derivations is the semantic composition. For
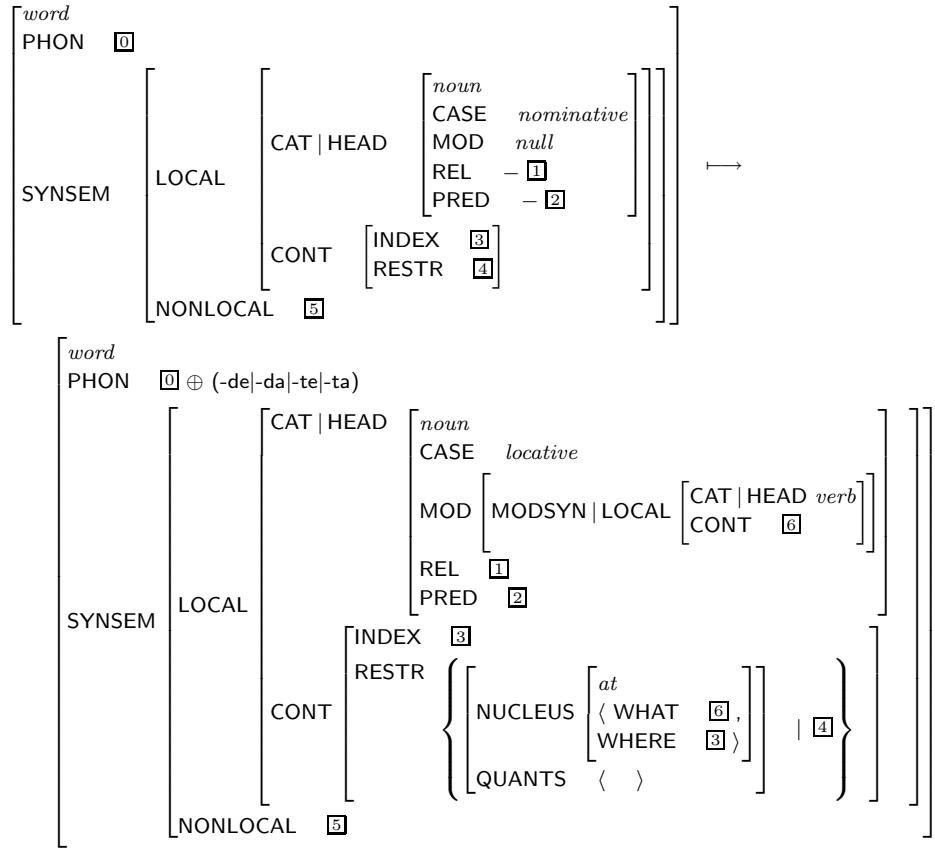
```
┌ word                                                                      ┐
│ PHON   [0]                                                                │
│         ┌        ┌                   ┌ noun                    ┐ ┐ ┐      │
│         │        │                   │ CASE   nominative       │ │ │      │
│         │        │ CAT | HEAD        │ MOD    null             │ │ │      │
│         │        │                   │ REL    − [1]            │ │ │      │
│ SYNSEM  │ LOCAL  │                   │ PRED   − [2]            │ │ │  ↦   │
│         │        │                   └                         ┘ │ │      │
│         │        │          ┌ INDEX   [3] ┐                      │ │      │
│         │        │ CONT     │ RESTR   [4] │                      │ │      │
│         │        └                                               ┘ │      │
│         │ NONLOCAL  [5]                                            │      │
└         └                                                         ┘       ┘

┌ word                                                                               ┐
│ PHON   [0] ⊕ (-de|-da|-te|-ta)                                                      │
│         ┌        ┌                   ┌ noun                                  ┐ ┐ ┐  │
│         │        │ CAT | HEAD        │ CASE   locative                       │ │ │  │
│         │        │                   │                ┌                  ┐   │ │ │  │
│         │        │                   │ MOD   │ MODSYN | LOCAL  ┌ CAT|HEAD verb ┐│  │ │ │  │
│         │        │                   │                │                 │ CONT    [6]   ││  │ │ │  │
│         │        │                   │                └                  ┘   │ │ │  │
│         │        │                   │ REL    [1]                            │ │ │  │
│         │        │                   │ PRED   [2]                            │ │ │  │
│ SYNSEM  │ LOCAL  │                   └                                        ┘ │ │  │
│         │        │          ┌ INDEX   [3]                                       ┐ │  │
│         │        │          │ RESTR    ⎧ ┌            ┌ at              ┐ ┐     ⎫ │ │  │
│         │        │ CONT     │          ⎨ │ NUCLEUS   ⟨ WHAT   [6] ,    │ │     ⎬ │ │  │
│         │        │          │          ⎩ │           │ WHERE  [3] ⟩    │ │  | [4] ⎭ │ │  │
│         │        │          │            │ QUANTS    ⟨   ⟩             │ │       │ │  │
│         │        │          └            └                            ┘ ┘       ┘ │  │
│         │ NONLOCAL  [5]                                                            │  │
└         └                                                                        ┘   ┘
```

Figure 3: Lexical rule for the locative case.

instance, the -CI morpheme (with allomorphs *-cı/-ci/-cu/-cü/-çı/-çi/-çu/-çü*) adds the meaning "doer/user of something" (7a), "seller/lover of something" (7b), or habitual (7c).

(7) a. *yol   -cu*
       road
       'traveller'

   b. *şeker -ci*
      candy
      'candy seller or lover'

   c. *sabah   -çı*
      morning
      'morning person'

Clearly, this ambiguity cannot be resolved without incorporating into lexical semantics a Qualia Structure a la Pustejovsky [?], or lexical semantic constraints [?]. We have been incorporating these types of constraints. Unfortunately, descriptive work on Turkish linguistics in this regard is very scarce, and there is no ontology such as Levin's [?]. Using features like [∓animate], [∓artifact], [∓container], and [∓period], one can define semantic fields for the derivational morphemes. We

expand the set of features as more lexical items are added to the lexicon. This is a very labour intensive task; the lack of a large-scale initiative on lexicography in the manner of LDOCE or COBUILD is hindering the efforts for automatic extraction of lexical knowledge from on-line resources.

Our strategy is to obtain complex forms derivationally if the semantic relation of the bound morpheme to its stem is fairly predictable. We use lexicalized forms when the meaning is not compositional. One such case is the denominal verb suffix -*le*, which is very productive but has no predictable meaning that can be derived from the lexical semantics of the stem.

Lexical Category Changes: As described in section 1, we model the nominal use of adjectives in Turkish by a single lexical item which may be interpreted as a term or a predicate by a lexical rule. There are other linguistic phenomena that are on the boundary of lexicon and syntax, which we opted to contain in the lexicon, e.g., non-referential objects, and valency change in the causatives. In the following, we briefly describe the lexical rules for them.
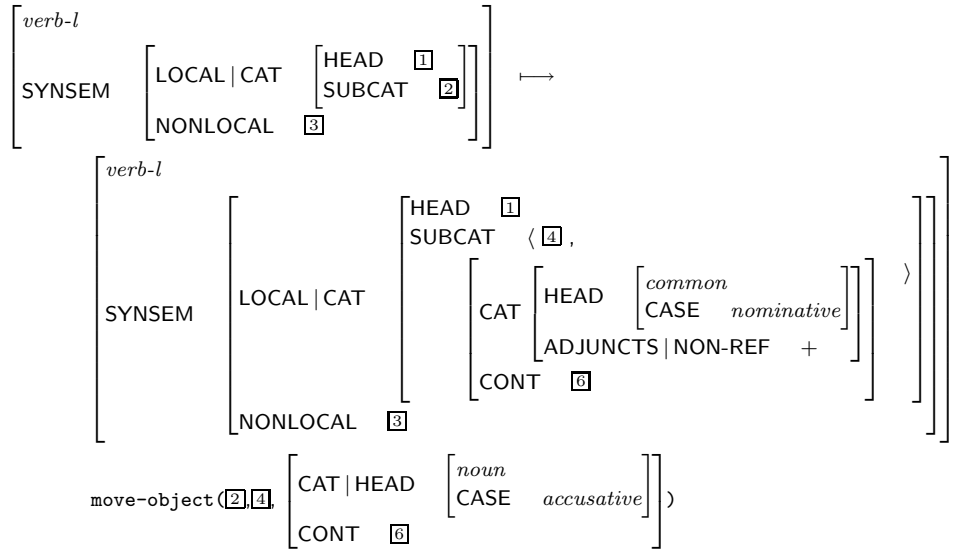
Case assignment is overt in Turkish, which allows for scrambling of the constituents. All six permutations of the SOV order are felicitous if the object NP is case marked (e.g., 8a and 8b). If the object is non-referential or indefinite (cf. 8a and 8c), it is not marked morphologically, which blocks scrambling, and the unmarked SOV order is used (cf. 8c and 8d).

(8) a. *Çocuk   kitab-ı   oku-du*
       child.NOM book-ACC(=object) read-TENSE.3SG
       'The child read the book.'

    b. *Kitab-ı çocuk oku-du*

    c. *Çocuk   kitap   oku-du*
       child.NOM book.ACC read-TENSE.3SG
       'the child read a book ($\cong$ the child did book-reading)'

    d. * *Kitap çocuk okudu*

Non-referential objects are not inflected, and they must occupy the immediately preverbal position. One way of dealing with nouns, then, is to keep two entries in the lexicon: one for unmarked form which may receive case marking and scramble, and one with lexically assigned case (accusative), which may not scramble. Our solution is to have a lexical rule that changes the subcategorization frames of verbs to handle cases where objects may be case-marked NPs or unmarked Ns. In the second case, the entity is marked indefinite and all scrambling is blocked by the lexical rule. Figure 4 shows the lexical rule in ALE notation (the rule is simplified for ease of exposition).

Causatives can be modelled in a similar vein. A causative suffix changes the subcategorization frame of the verb by adding one more argument and changing the grammatical constraints on the other arguments. For instance, the new argument becomes the subject (causer), and the old subject (agent) is demoted down the grammatical hierarchy [**?**] to direct object or indirect object, depending on the valency of the verb:

(9) a. *Can arkadaş-ı-nı   çağır-dı*
              friend-POSS-ACC call-TENSE.3SG
       'Can called his friend.'

$$
\begin{bmatrix}
\textit{verb-l} \\[4pt]
\text{SYNSEM} \quad
\begin{bmatrix}
\text{LOCAL}\,|\,\text{CAT} &
\begin{bmatrix}
\text{HEAD} & \boxed{1} \\
\text{SUBCAT} & \boxed{2}
\end{bmatrix} \\[8pt]
\text{NONLOCAL} & \boxed{3}
\end{bmatrix}
\end{bmatrix}
\longmapsto
$$

$$
\begin{bmatrix}
\textit{verb-l} \\[6pt]
\text{SYNSEM}
\begin{bmatrix}
\text{LOCAL}\,|\,\text{CAT}
\begin{bmatrix}
\text{HEAD} & \boxed{1} \\
\text{SUBCAT} & \langle\, \boxed{4}, \\
& \quad
\begin{bmatrix}
\text{CAT}
\begin{bmatrix}
\text{HEAD} &
\begin{bmatrix}
\textit{common} \\
\text{CASE} & \textit{nominative}
\end{bmatrix} \\[6pt]
\text{ADJUNCTS}\,|\,\text{NON-REF} & +
\end{bmatrix} \\[6pt]
\text{CONT} & \boxed{6}
\end{bmatrix}
\,\rangle
\end{bmatrix} \\[6pt]
\text{NONLOCAL} \quad \boxed{3}
\end{bmatrix}
\end{bmatrix},
$$

$$
\texttt{move-object}(\boxed{2},\boxed{4},
\begin{bmatrix}
\text{CAT}\,|\,\text{HEAD} &
\begin{bmatrix}
\textit{noun} \\
\text{CASE} & \textit{accusative}
\end{bmatrix} \\[6pt]
\text{CONT} & \boxed{6}
\end{bmatrix})
$$

Where `move-object` *is a definite clause which deletes the accusative object from the* SUBCAT *structure in first argument and return resulting structure and accusative object in second and third argument respectively.*

Figure 4: Lexical rule for non-referential objects.

b. *Mehmet Can-a   arkadaş-ı-nı   çağır-t-tı*
    Can-DAT friend-POSS-ACC call-CAUS-TENSE.3SG
'Mehmet had Can call his friend.'

Morphophonemic rules: The rules for inflectional and derivational morphology might also take into account the archiphonemes that are not marked for certain features. For instance, the locative case marker has allomorphs *-de/-da/-te/-ta*. They may be represented uniquely by two metaphonemes -DA where D is a dental stop unmarked for voice and A is a low unround vowel unmarked for backness/frontness. Vowel harmony and voicing constraints[4] determine their surface realization during morphological composition. These kinds of rules are not lexical rules per se since they do not operate on lexical properties of the words. In our model, they are embedded in lexical rules for inflections and derivations.

4   Conclusion

For a language with rich morphology, lexical rules can be used for controlled generation of surface forms. Inflections and derivations can be seen as word-based (local) operations on the root, and thus be modelled as lexical rules. Phonological alternations in stems can be embedded in the rules as well. Grammatical role changes, type constraints on word subtypes, and noun to NP promotions (as in non-referential objects) control the proliferation of lexical entries. Semantic contribution of inflections seems to be morpheme specific: All derivations take part in semantic composition, but some inflections (such as case and causatives) contribute semantically as well. Most inflections (e.g., person and number markers), however, have grammatical functions only. This is not to say

---

[4]cf. [**?**, **?**] for a description of these processes. [**?**] is the original work on Turkish that combines finite state morphotactics with morphophonemic alternations.

they do not have a semantic form, just that in many cases the form is that of identity. Productive use of derivations is limited by the predictability of the semantic relation of the stem to the affix.

We have been testing our lexicon design as part of an HPSG grammar for Turkish [**?**]. The grammar development environment, ALE, had to be modified to allow run-time evaluation of lexical rules. Compiling out the lexical rules seems to be impractical, since generating every possible form for a large lexicon of roots causes exponential growth in the lexicon. Compilation of all surface forms for a lexicon of only 40 root forms produces around 2800 entries, and takes about 8 minutes on a Sun Sparcstation 10. Run-time execution of rules puts the burden on parsing or generation. We believe that as the lexicons of NLP systems become more comprehensive and open-ended, the trade-off will be resolved in favour of using the lexical rules on demand at the expense of slower performance.